Overfeat: Classification, Localization

by Gradient Ascent: Ding Ding Wei, John Manning, Oliver Kanders
CSCI 1470 (Deep Learning), Department of Computer Science, Brown University

Introduction

Object detection remains a significant point of focus for tackling greater problems that involve the comprehension and analysis of images and videos. The concepts of Classification and Localization of these images prepares a new frontier for object prediction: labeling the object, describing its category, and then providing the location of the object within bounding boxes. The OverFeat model achieves state-of-the-art performance on object detection tasks, demonstrating the effectiveness of CNNs for object detection. Since then, numerous variants of CNN-based object detection models have been proposed, each with its own unique set of strengths and weaknesses. The model is based on the paper *Overfeat: Integrated Recognition, Localization and Detection Using Convolutional Networks*.

Layer	1	2	3	4	5	6	7	Output 8
Stage	conv + max	conv + max	conv	conv	conv + max	full	full	full
# channels	32	64	128	128	256	385	512	3
Filter size	11x11	5x5	3x3	3x3	3x3	-	-	-
Conv. stride	4x4	1x1	1x1	1x1	1x1	-	-	-
Pooling size	2x2	2x2	-	-	2x2	-	-	-
Pooling stride	2x2	2x2	-	-	2x2	-	-	-
Zero-Padding size	-	-	1x1x1x1	1x1x1x1	1x1x1x1	-	-	-
Spatial input size	231x231	24x24	12x12	12x12	12x12	6x6	1x1	1x1

Figure 1. Model architecture detailing the reduced Overfeat Extractor and Classifying Layers

Methodology

The model comprises of three techniques for object detection, each a subgroup more complex than the next: Classification, Localization, and Detection. A basic Overfeat feature extractor was deployed, trained simultaneously with the Classification and Localization classes (a regressor on bounding boxes that would use the weights of the feature extractor to predict the locations of the labels). Categorical Cross-Entropy Loss and Mean Squared Loss assisted in minimizing optimization for the two tasks. The Classification module also consisted of a sliding window technique of size (2,2), which allows for different max poolings at different scales across the image, capturing greater information about objects present in the image. Once trained, the bounding box predictions were then merged across other bounding boxes based on the paper's criterion of a match score over 50% for intersection over union (IoU). Detection training was omitted.

Data

The dataset used in the model paper was the Microsoft COCO 2017 Object Detection dataset that contained over 120,000 images, 6,600 of which we examined for training. We specifically training the model to classify and localize three instances car, traffic light, stop signs

Training

Overfeat + Classifier = Classification

- We train this combination for 20 epochs using Adam optimizer for 20 epochs, with learning rate 0.001 and 0.0005

Overfeat + regression = bounding boxes

- With the trained feature extractor, we train the regression box for 20 epochs with learning rate 0.0005

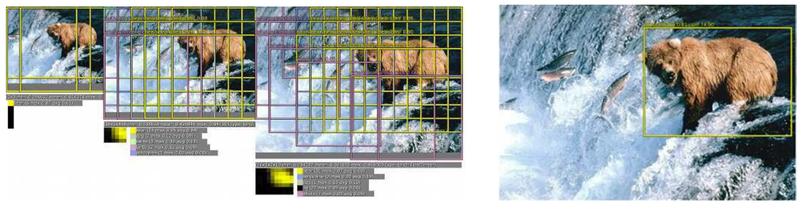

Figure 2. (left) Classifier of the image outputs a confidence for each location, then sliding window enhances prediction.
(right) Predictions are combined

Results

Our best training results, thus far, remain at a loss of 439 over 20 epochs. We were careful to not overfit the data, and leave space for the combination of bounding box predictions to occur.

The current regression model has a loss of 490 on the test data, and the classification model has ~93% accuracy on the test data

REGRESSION: Epoch	Training Loss
15	623.3777465820312
16	461.8003234863281
17	454.8875732421875
18	459.6468811035156
19	439.5505676269531

Figure 3. Model training metrics.

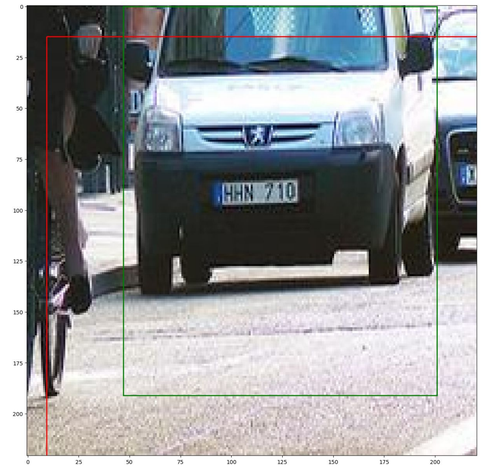

Figure 4. (red) Predicted Bounding Box
(green) Ground Truth Bounding Box

Discussion

The model performed extremely well in general classification, reaching thresholds of above 93% in label prediction for testing. The regression component remained difficult in aptly scaling these bounding boxes to capture the information for multi-classification. While our model performed the best for single class regression such as car, per class regression, where regressions of bounding boxes ranged for car, traffic light, and stop signs were more problematic to manage. In future work, we would hope to focus on creating many scales for the image, as noted in the paper, which would exponentially increase the amount of bounding boxes to work with and enhance our prediction for multi-classification. We would also look to deploy a custom detection training component, where false positives would be rejected, thus furthering our predictive powers. Additionally, the authors of the paper recommend expanding back-propagation for localization throughout the entire network (we currently freeze the Overfeat training weights).

Object detection remains a fundamental use-case for computer vision and carries a diverse range of applications in various fields: the Overfeat model proposes an exciting architecture for this detection, and look forward to future iterations on the design.

Reference

Sermanet, Pierre & Eigen, David & Zhang, Xiang & Mathieu, Michael & Fergus, Rob & Lecun, Yann. (2013). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. International Conference on Learning Representations (ICLR) (Banff).